Dissecting the Learning Behaviors in Hacker Forums

Alex Tsang Xiong Zhang Wei Thoo Yue
Department of Information Systems, City University of Hong Kong, Hong Kong
inuki.zx@gmail.com, xionzhang3@student.cityu.edu.hk, Wei.T.Yue@cityu.edu.hk
Michael Chau
School of Business, The University of Hong Kong, Hong Kong, mchau@business.hku.hk

Abstract: It is a consensus that there exists learning barriers to become a skilled hacker. Hack forums are believed to be indispensable venues where to overcome these barriers. However, this belief has not been formally justified in academia. By studying post content, we find evidence of different posting behaviors in hack forums. One interesting result is that there are indeed some learning behaviors exhibited by the changes in posting behaviors of forum users. Our results show that information exchange with regard to security knowledge is occurring in public forums, which contradicts to the common perception of hackers' mystique.

Keyword: text mining, hack forum, learning behavior, supervised classification.

1. Introduction

Recent attacks by *Anonymous* groups have once again put hackers back into news headlines. Causing loss of $388 billion in 2011 [1], attacks in information systems have become serious threats to organizations. However, hackers' identity has always been elusive to the public [2]. The common perception is, to evolve into hackers with comprehensive and broad skills, people need to learn and get trained through mysterious hacker curriculum [3]. Hackers usually tend to interact and share knowledge with others in cyberspace [4]. Online hack forums are one type of venues where this kind of information exchange could take place [5]. However, so far, there aren't any formal academic study about hackers' learning behaviors in online hack forums. This paper aims to fill this gap. Specifically, we focus on distinctive information flow amongst forum users based on analysis of their post content. Our results show that hack forums are indeed a de facto learning community where newbie hackers can ascend the membership ladder to become skilled hackers [6]. We capture the structural changes in post content which represent the changes of learning behaviors in forums. Also we find evidence that some hackers exhibit the learning characteristics of a decreasing trend of information inflows and an increasing trend of information outflows over time.

2. Data Collection

Data is collected from a popular hack forum in the range of 6 years. Both hack-related and non-hack-related topics aggregate into dozens of sub-boards, of which 19 hack-related sub-boards focus on hacking topics and hacking technologies, in this forum. Post-centric information i.e., title, content, date, view count and reply count, as well as user-centric information i.e., users' ID, level, join date and post count, are collected in our dataset. This dataset includes 76738 users initiating 354826 threads, with a total of 2969171 posts.

3. Model and Implementation

With the dataset collected, we first examine post content to see whether they demonstrate any patterns of information flow. Support Vector Machine (SVM) is used to automate the post

content classification process [7]. Due to the complexity of the post contents, we restrict our analysis to headers in a particular thread. The changes of posts' category are then studied based on the classified post content.

Post Category

We focus on two general categories of posts based on the information flow of post content:

Category 1 (Information inflow): including questions, doubts, requests, advice seeking, and anything else reflecting the user's requirement for information.

Category 2 (Information outflow): including answers, tutorials, teachings, troubleshooting, guidelines, demonstrations, showoffs and anything reflecting offers of information.

The defined categories serve as the basis for a supervised learning for post classification. We adopt Joachim's SVM^{light} and make use of its one-versus-all property of binary classification to classify post category [8]. Due to the large amount of data, we focus on thread header posts.

Classification Features

A total of 9 attributes of data concerning each post are considered as in Table 1. These are the useful knowledge we can have about posts and users in our dataset.

Fields for features	Description
postTitle	Title of the entire thread
postContent	Content of the first post in the thread, i.e., the header post
imageCount	Number of images appearing in the content
linkCount	Number of hyperlinks appearing in the content
postView	Number of times this post has been viewed
postReply	Number of replies of this post
userLevel	User's level when posting the post
userPostNumber	Number of posts by this user when posting the post
contentLength	Word count of post content

Table 1: Post Features

Keyword Extraction

To extract meaningful keywords from post title and content, the N-gram stemming procedure is performed on *postTitle* and *postContent* respectively, where N = {1, 2, 3, 4, 5}. In this procedure, each text is processed through a Porter stemming procedure [9]. All of the stemmed words are reassembled as a space-delimited text again for keyword extraction. A keyword for title is recognized if it appears at least twice in all the post titles, and a keyword for content is recognized if it appears at least 5 times in all the post contents. To summarize, 228 keywords for titles and 2254 keywords for contents are extracted. In each post category, keywords are assigned consecutive feature IDs. Finally, for each keyword, a global IDF (*Inverse Document Frequency*) value is calculated for subsequent use, with $idf(T, D) = log \frac{|D|}{|\{d \in D : t \in d\}|}$.

Model Creation

One training set and one test set are created for the SVM^{light} classifier to generate models for post category classification. The training set is composed of a random selection of 11 posts from each of the 19 sub-boards and the test set is composed of a random selection of another 6 posts from each of the same sub-board. The number of selected posts is proportional to the size of hack

forum dataset. We manually classify the posts in both sets as in Table 2.

Each post in the training set is examined to identify any keywords beforehand, and each identified keyword is represented by a TF*IDF value with $tf * idf(t, d, D) = tf(t, d) \times idf(t, D)$, in which the IDF value is retrieved from the prior calculation results.

Each TF*IDF value for a keyword found in a post is included as a feature prefixed with its ID. Finally, the training set is used to train SVM^{light} classifier. With appropriate parameter tunings, we obtain the models for both post categories, with higher than 80% for all accuracy, precision and recall values on the test set as in Table 3.

	Training Set	Test Set
Category 1	93	58
Category 2	92	53
Others	24	3
Total	209	114

Table 2: *Training Set and Test Set*

	Post Category 1	Post Category 2
Accuracy	84.21%	84.21%
Precision	80.65%	80.33%
Recall	89.29%	89.09%

Table 3: *Classification Rate in Test Set*

Post Classification

The generated models are used to classify all the posts in the corpus. Since it is a one-versus-all binary classification, any posts simultaneously classified into both categories will be classified into the category with larger prediction rate. Based on the classified posts, we examine the changing trends of posting behaviors in these 2 post categories for each user, and further reveal users' behavioral pattern.

Post Evolution

Intuitively, the possible variation trends (i.e., increase or decrease) in both categories constitute a total of 4 patterns of user behaviors in the forum. A user with confirmed variation trends in both categories can be grouped into one of these 4 patterns. We then perform Linear Regression on the post quantities in both categories against the fixed quantity of posts posted in a sequential manner (We take an increment of 5 posts as a proxy for time point for a user).

Among 76738 users posting all the 354286 posts, we pick the top 1000 users in terms of post count, those who had cumulatively contributed 65279 posts. In this group, the average post count per user is 65.3, with the largest (or smallest) post count per user is 599(or 39), compared with an average of 3.8 posts per user outside this group. By setting 5 posts as a time point, we guarantee a minimum of 8 data points as a sensible regression for each user. Regression results whose standard error is larger than the average are also excluded.

4. Analyses and Results

With the exclusion of unreliable regression results, we have 318 users remaining with clear variation trends of both post categories. By examining the slope of the regression model in 2 post categories for each user, we can justify users' behavioral pattern. Numbers of users under each of these 4 posting behavioral patterns reveal their proportional relationships, as in Table 4.

Among the 318 users, we find that a fair number of users (45.6%) exhibited typical patterns of learning behavior with decreasing requests for information inflow and increasing information outflow. All users who exhibit this type of information flow are grouped and their collective behaviors are shown in Figure 1. As we set the incremental unit to 5 posts, each scale on the x-axis captures every additional 5 posts. Essentially, the x-axis represents the different time stamp at which the users participated in the forum (i.e., first 5 posts, time point t =5, second 5 posts , t =10) and reveals the categorical distribution for every 5 posts. In Figure 1, when t =10, for post 5-10, we have an average Cat_1 =2.484, Cat_2 =1.942. The total is less than 5 because some posts are classified out of these 2 categories. By examining on average how many posts belong to Category 1 and Category 2 respectively (y-axis) for every 5 posts (x-axis), we see that Pattern 1 has a relatively steep slope, which indicates a robust learning activity is taking place.

	Category_1 (Avg. value)			Category_2 (Avg. value)		
Total: 318	Slope	Intercept	R-squared	Slope	Intercept	R-squared
Pattern 1:145	< 0 (-0.153)	4.014	0.298	>= 0 (0.086)	1.082	0.152
Pattern 2: 86	>= 0 (0.082)	1.978	0.155	< 0 (-0.120)	2.984	0.248
Pattern 3: 81	< 0 (-0.060)	2.956	0.077	< 0 (-0.058)	2.269	0.100
Pattern 4: 6	>= 0 (0.009)	2.352	0.002	>= 0 (0.027)	2.228	0.026

Table 4: *Average values for user behavioral patterns*

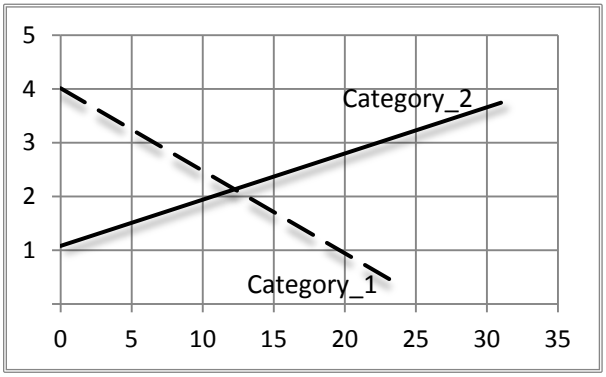

Figure 1: Average User Posting Behavior: Pattern 1

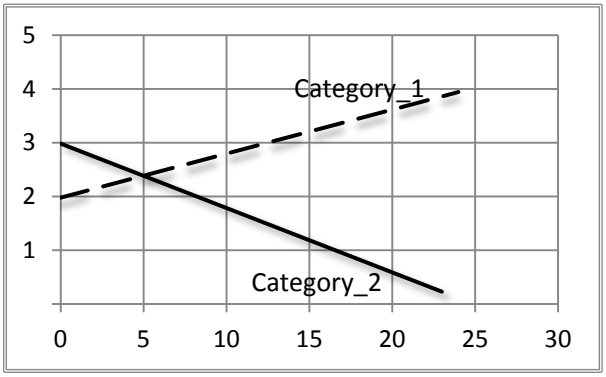

Figure 2: Average User Posting Behavior: Pattern 2

We also see that 27% of the users in the dataset exhibit reverse behavior pattern with an increasing information inflow and a decreasing information outflow as in Figure 2. They tend to ask more questions and share less information as they raised more and more topics in the forum. While dedicated to "knowing more" and "learning more", they pay less attention to the promotion of their virtual status by demonstrating their knowledge.

Meanwhile, another 25.5% of the users in our dataset reveal a decreasing information inflow and outflow. The flatter average variation trends of information flow indicate that these users have either lost interest in both asking and sharing or had probably engaged in other activities. The remaining 1.9% of the users shows increasing information inflow and outflow, and hence is deemed to be insignificant because of their low quantity and the near-to-flat slope. Due to page limit, we remove figures for these two groups of users.

Through examining the 145 users in Pattern 1, we find that, for the 128 users who had ever posted at least 5 posts in Category2, there is an increase in the replies to these posts, which rises from an average of 9.04 replies to the first 5 posts, to an average of 11.17 replies to the last 5

posts. This phenomenon indicates that the actively learning users are actually gaining more recognition in the forum, signaling a gaining of a higher status in the community.

5. Conclusion and Future Work

Internet is an excellent medium for sharing information. Anecdotal evidence has shown that Internet has provided the environment to facilitate exchange of hack knowledge [10]. Our results show that there is a fair amount of learning behaviors taking place in hack forum, which would transform user behaviors. The result is both expected and surprising. On one hand, we expect the learning activities to occur, but on the other hand, the learning patterns are perhaps not as evident in a public forum. The result challenges the common perception of the mystique of hacker groups. One explanation is that given the fact that hacker activities are decoupled from actual attacks, there is really no hindrance in sharing information. Furthermore, the observed behaviors could be tied to the white-hat hackers, many of whom are IT security professionals.

We can extend our current work in many directions. First, we can further study the interactions amongst users in the forum. Social network analysis could be performed to further observe different interaction groups. Second, the nature of the posts in relation to actual attacks could be further analyzed. For instance, we can link post discussions with actual vulnerabilities reported in a vulnerability database. We can also study the relationship of hacker posts to a secondary attack data source. Besides, we can also extend this paper by studying what kind of knowledge can be quickly learned by hackers.

Reference

[1] "Norton Study Calculates Cost of Global Cybercrime:$114 billion Annually," 7 September 2011. [Online]. Available: http://www.symantec.com/about/news/release/article.jsp?prid=20110907_02.

[2] T. E. Heinzen and L. M. Picciano, "Ilk Hunting: Newbies, Cyberpunks, Coders and the Search for Elusive, Ego-Twisted, Talented Computer Hackers," in *International Handbook on Giftedness*, Springer, 2009, pp. 809-823.

[3] S. Bratus, "Hacker Curriculum: How Hackers Learn Networking," *IEEE Distributed Systems Online*, 2007.

[4] D. E. Denning, "Hacker Ethics," [Online]. Available: http://www.southernct.edu/organizations/rccs/oldsite/resources/research/security/denning02/introduction.html#introduction

[5] R. DuFour, "What Is a Professional Learning Community?," May 2004. [Online]. Available: http://www.ascd.org/publications/educational-leadership/may04/vol61/num08/What-Is-a-Professional-Learning-Community%C2%A2.aspx.

[6] A. D. Smith and W. T. Rupp, "Issues in cybersecurity:understanding the potential risks associated with hackers/crackers," *Information Management & Computer Security*, pp. 178-183, 2002.

[7] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd Annual International ACM SIGIR Conference of Research and Development in Information Retrieval*, Berkeley, CA, USA, 1999.

[8] T. Joachims, Learning to Classify Text Using Support Vector Machines-Methods, Theory and Algorithms, Kluwer Academic Publishers, 2002.

[9] M. F. Porter, "An Algorithm for Suffix Stripping," *Program*, pp. 130-137, 1980.

[10] P. Olson, We Are Anonymous: Inside the Hacker World of LulzSec, Anonymous, and the Global Cyber Insurgency, Little, Brown & Company, 2012.